Analysis of Cross Domain Sentiment Techniques

Ashwini Save
Computer Engineering Department
D. J. Sanghvi College of Engg,
Mumbai, India
ashwini.save@gmail.com

Narendra Shekokar
Computer Engineering Department
D. J. Sanghvi College of Engg,
Mumbai, India
narendra.shekokar@djsce.ac.in

***Abstract* – Sentiment analysis is the analysis of opinions on certain entities. Due to the wide applications sentiment analysis has been widely researched. But due to the usage of sentiment analysis in wide number of domains there are multifaceted issues and problems which has kept the researchers on their toes. One such problem is usage of training data of one domain for classification and also lack of annotated data for training of the classifier so that it can perform prediction accurately. This is where cross domain sentiment analysis comes in. The paper performs a brief survey and analysis of cross domain sentiment analysis approaches and algorithms.**

Keywords – Sentiment Analysis, Cross Domain Sentiment Analysis, Machine learning, Cross Domain Adaptation.

I. INTRODUCTION

The number of internet users is growing at an unprecedented rate all over the globe. At the start of year 2017 internet users around the world amounted to around 3.7 Billion, with Asia accounting to almost half of it at 49.6% [1]. By the end of 2015 India had more than 327 million internet users, with an estimate of it reaching in excess of 600 million by the year 2020 [2]. By June 2016 India already had around 462 million internet users [3]. If the internet user base is growing at around 10% then India is witnessing a 30% plus increase in internet traffic year on year.

This shows that as more and more people are embracing the internet, many of them are using the platform in either expressing their opinions or using these customers opinion to make informed decisions.

This use of internet in expression of opinions has given rise to humongous amount of data. When this data is effectively used by commercial enterprises to find customers buying patterns and other information this data is called as commercial databases, which are witnessing a rise at an unprecedented rates.

Data by itself is useless, important information has to be recovered from it. Clive Humby, UK Mathematician and architect of Tesco's Clubcard, very aptly puts it "Data is the new oil. It's valuable. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value" [4].

This is where Data Mining comes in. In data mining hidden information is extracted from large databases for prediction of some kind. And, one of the important specialisation of data mining is opinion mining or Sentiment Analysis. In sentiment analysis people's opinions or sentiments towards some entity is analysed for some specific purpose.

The sentiment analysis has been a field which has seen multitude of research in the recent times. There are many reasons for it. If done properly then sentiment analysis would prove to be very useful for the business and customers alike. If Sentiment analysis is carried out with precision it proves to be of great advantage as it can provide insights which can improve customer service, improve success of the campaign, determine marketing strategy, improve product messaging, , test business Key Performance Indicator's (KPIs) and also generate leads for better business [5].

As far as shoppers are concerned if there is a system which can analyse the sentiments properly and provide good results then that would lead to buying of good products according to ones requirements. Sentiment analysis has many applications, like in review-based websites, in business intelligence, in online commerce, etc. Due to these reasons the field of sentiment analysis has been highly researched.

A. Challenges in Sentiment Analysis and Cross Domain Sentiment Analysis

1. Domain Specific Training Data : Sentiment analysis will be able to provide proper insights when it has been properly trained. For adequate training of the sentiment analysis system a respectable amount of annotated or labelled training data is required; the labelling of the training data is a manual process.

2. Exponential Increase in number of products and its categories : In recent history the number of products and its categories has grown with leaps and bounds. This has posed a great challenge for proper analysis of sentiment. This is because it is difficult to have a respectable amount of labelled training data for all the different products categories.

978-1-5386-2361-9/17/$31.00 ©2017 IEEE

3. Lack of Annotated Data : There is scarcity of annotated training data to carryout proper training of the classifiers. As there is a lack of annotated training data at our disposal the sentiment analysis system is not trained properly and consequently the results produced by the system will not be accurate.

4. Same word different meaning in different domains : It is seen that a same word might have different meaning in different domain along with different polarity. So, a classifier trained on two domains where there exists at least a word having different meaning and polarity in these two domains the accuracy of the classifier is compromised.

This is where cross domain sentiment analysis comes in. Because of these and other reasons cross domain sentiment analysis and its techniques is being researched very extensively in the recent times.

II. CROSS DOMAIN SENTIMENT ANALYSIS

Sentiment analysis, in most cases, is carried out for specific domains to get better results. Here, the sentiment analysis classifier is usually trained on the features which are domain specific; here the domain refers to Books, Hotel, Movies, Mobiles, Microwave Owens, etc. So, a classifier trained on a specific domain will fail when it is fed with sentiments expressed on a different domain. Hence annotation of respectable amount of data for proper training is required. This annotation of the data is a manual process. And, considering the fact that the number of domains keeps on increasing it becomes very difficult, time consuming and costly to annotate the training data.

Due to these reasons an approach is required which would be able to estimate and produce a results even when the classifier is trained on different domain. This approach is called as Cross Domain Sentiment Analysis.

If a classifier is trained on a specific domain and is used to predict for a different domain data it results in very poor performance. So, in cross domain sentiment analysis the classifier tries, as much as possible, to adapt itself to the unknown target domain. This adaptation of the classifier is the most important aspect of cross domain sentiment analysis.

In the recent times many techniques have been put forth to achieve proper cross domain adaptation or cross domain sentiment analysis. One of the important technique in achieving domain adaptation is the usage of pivots. Pivots are those words and features from both the source and the target domain which represent the same sentiment [6].

John Blitzer, et. al. [6] propose a method for selection of pivots. The paper sues the structural correspondence learning technique for pivots selection. The proposed method works in two steps for the selection of the pivots. First the frequency of the words are found out from the training data, the words with maximum frequency can be called as the pivot of the source domain [7]. Further to increase the accuracy of the domain adaptation paper [7] proposes to provide with a labelled data. Now the pivots are selected based on two criteria, the words with the highest frequency and their mutual information with the labelled source data. On the down side the paper requires labelled source pivots which is manual process.

On the other hand Danushka Bollegala, et. al. [8] skips the whole process of pivot selection. The paper assumes that the pivots have already be given. Once the pivots of source and target domain are given the paper puts forwards three rules. First, the pivots of source and target domains should be as close to each other as possible. Second, the distance between same sentiment words (positive or negative) should be as less as possible and the distance between words of two different classes should be as greater as possible. The reason for doing is that, it decreases the margin of error. Third, the paper proposes to maintain internal relationships between the pivots and the non-pivots in both the source and the target domain. These are the three objective function which the papers tries to optimize to get better domain adaptation.

The paper produces good accuracy but the fundamental problem is that the paper assumes that ‘m’ number of pivots have already been given to the system.

Zhen Hai, et. al. [9] proposes a joint topic modelling approach. The paper provides a supervised learning approach for achieving cross domain sentiment adaptation. Here form the target domain opinion pairs are selected and with the help of labelled source data aspects terms are generated for the target domain. With the help of these generated aspect terms the classifier determines to which of the source domains the target domain adapts appropriately. To be sure of the results produced by the classifier is accurate and to verify the results predicted by the classifier the proposed work makes use of the star rating given by the reviewer.

The papers claims to produce an accuracy of around 80% for the taken dataset. But as the proposed system uses supervised learning approach it cannot be said as a pure cross domain approach. This is because, as it is a purely supervised learning approach the classifier has to be trained on all possible target domains as well. If it is not trained on all the different possible target domains then the system will fail.

Shenghua Liu, et. al. [10] works exclusively for the data produced on Twitter. The paper proposes a technique to identify the topic or the domain of the target data and further performs sentiment classification. Here first the system divides a particular tweet into text and non-text. If the non-text characters contains emoticons then appropriate polarity is given. The non-text characters also contains time stamps of the tweet as well. The paper assumes that if the time stamps of the tweets are similar, that is if the time difference between the tweets are very less, then it can be said that both the tweets are of the same topic or domain. Also, in the tweets ‘@’ is used to address someone. The system proposes to develop a hierarchy of ‘@’, this will help in identifying the source of the topic and possibly the topic of discussion as well.

For the text (words), the paper again divides it into two parts global words and non-global words. Global words are those words which do not change their polarity in any domain, like good, bad, etc. polarity of all the global words are calculated. To find the domain, the non-global words are used by the paper and this is achieved with the help of transduction method.

In a general supervised learning the classifier creates a general rules from the labelled training data and then applies these rules on the test cases. On the other hand in the transduction method rather than creating general rules, some kind of relation is found out between the test cases and the labelled data individually.

Kun-Hu Luo, et. al. [11] proposes a simple method for achieving of cross domain sentiment analysis. The paper proposes to use only emotional words for sentiment classification. According to the paper emotional words are those words which do not change their polarity (like global words) across domains. The paper proposes to extract all the emotional words from the target domain and then classification is performed.

To solve the cross domain resignation problem Yuewei Lin, et. al. [12] in traduces a new method of developing joint subspaces. A subspace is a cluster of data points having similar attributes and characteristics. These subspaces are created one for each class. The proposed system has been implemented in six steps. In the first step a subspace is created for the source domain. Anchors are identified and an anchor subspace is created. Anchors are those words which carry the information of target exclusive characteristics. Also, the paper says that the anchor subspaces should be compact.

Then the anchor subspaces are assigned label individually. For proper assignment of labels to the anchor subspaces two things are considered, relation between source and target subspaces and relation within anchor subspaces. That is, to assign some class to each anchor subspace these two things are considered to find similarity between anchor subspace and the source subspaces.

The fourth step is to construct the joint subspaces. Maximum number of overlapping means that is how closely the two domains complement each other. And then the classifier is trained on this constructed joint subspaces.

In this paper the source and target domains pairs will have to be identified beforehand this is a major drawback of this system. The system fails if the target domain is different than the domain used for the creation of the joint subspaces.

Rather than using joint subspaces for the training of the classifier as in [12], Shuang Li, et. al. [13] proposes to train the classifier after reweighting of the data points. The paper proposes to utilise labelled instances from the source domain and unlabelled data of the target domain to predict the unknown labels of the target domain.

The paper assumes that when the source domain data points and target domain data points are plotted on a hyperplane the source classifier will give more accurate predictions on the target instances which are closest to the source domain.

Here the source and the target domain are plotted on same hyperplane. Then with the help of the domain separator the distance between the plotted data points and the domain separator is found out. Then the data points which are closest to the domain separator is assigned the highest weight. As the data points move away from the domain separator the weight assigned decreases. This reweighted data points are used for the training of the classifier.

For the prediction of the unknown target data point two things are done. First with the help of the reweighted source data point the reweighted target data point is predicted. Then with the help of the predicted target data point the unknown data point is predicted by using label propagation.

The paper proposes a novel and relevant way of achieving cross domain sentiment analysis. Reweighting scheme proposed by the paper is quite an effective way of training classifier for domain adaptation. But, as with the previous paper even here the source and target domains are a predefined pairs have to be identified beforehand. And, data points belonging to any other target domain will not be predicted accurately.

Debora Nozza, et. al. [14] propose the usage of deep learning concept for achieving cross domain adaptation. With help of deep learning concept the feature values of the source domain is transformed. If after transformation the feature values are similar to the features values of the target domain then it can be said that the corresponding features are similar as well.

Whereas, Wenjie Zhang, et. al. [15] proposes a system to find out the similarity between similar words. The paper works with the assumption that finding or distinguishing identical concepts across domains is difficult. So, the extracted words from the source domains and target domains are divided into two types, similar words and distinct words. Then similarity between the similar words is found out. if they are similar then the features of the source domain can be approximated to the feature of the target domain. Here the major drawback is that the source and target domain has to be defined beforehand. Also, the process of dividing the words into similar and distinct words is a difficult task.

Danushka Bollegala, et. al. [16] propose to develop and use a sentiment sensitive thesaurus for sentiment classification. The thesaurus is automatically developed with the help of semi-supervised learning approach. It makes use of the labelled and unlabelled source domain data and unlabelled target domain for training and testing purposes.

The paper considers only binary labelling, that is positive or negative. Also, only the adjectives of the source domain is taken for labelling; unigrams and bi-grams are taken for labelling. These unigrams and bi-grams are labelled as positive or negative. Then the system works in two steps. In

the first step with the help of the labelled source data relatedness is found out between unlabelled and labelled source domain data. With the help of the relatedness a Feature expansion is carried out for the target domain which is unlabelled. In feature expansion the feature vectors are augments with additional related features with the help of the thesaurus prepared. The additional related features are added by using a ranking mechanism proposed in the paper.

The major drawback of the proposed system is that it uses pointwise mutual information very extensively and hence the complexity of the system increases. Also, for the preparation of an exhaustive thesaurus a large number of labelled training data will be required which is detrimental and expensive.

Working on the paper [16] K. Aarthi et. al. [17] proposes to develop a Cross-Domain sentiment classification system to deal with the problem of feature mismatch. It makes use of labeled data from source domain and unlabeled data from source and the target domains. Relationships between different words in thesaurus is improved by multiplying a discounting factor with the point wise mutual information. Then the proposed system extends feature vectors by using the created thesaurus. Using which a binary classifier is trained from the source domain labeled reviews to predict positive and negative sentiment in reviews.

The proposed method only performs binary classification i.e. the reviews are classified as either being positive and negative. But, it can be extended to perform Multiclass classification of positive, negative and neutral reviews. It can also be extended to overcome the problem of word polymesy in cross-domains.

J. Karthikeyan et. al. [18] identified that in the paper [16] finding the multiple relatedness was a complex task. So, they proposed a system for cross-domain sentiment classification that uses L1/2 regularized Classification model at the feature expansion step. The paper claims that by using by L1/2 classifier the prediction accuracy of cross-domain sentiment classification improved. Also, the paper claims that it outperforms prediction accuracy of L1 classification.

P.Sanju and T. T. Mirnalinee [19] further tries to improve the thesaurus proposed in [20] by taking into account Wiktionary apart from the source and target domain data. Even though the paper proposes to develop a larger thesaurus compared to the original paper the computational cost of developing and using such large thesaurus is an important factor which cannot be overlooked.

In a practical scenario the feature distribution of target domain and source domain may vary drastically. In a worst case scenario the domain adaptation would fails if the feature distribution is completely different. In such cases. To deal with this feature distribution divergence across domains and heterogeneous feature representations of different domains, Min Xiao and Yuhong Guo [20] propose a feature space independent semi-supervised kernel matching method for domain adaptation.

thesaurus is prepared containing labelled and related unlabelled data of the source domain.

The system learns a prediction function on the labelled source data. And, the mapping the target data points similar to the source data points is done by matching the target kernel matrix to a submatrix of the source kernel matrix based on a Hilbert Schmidt Independence Criterion.

Fangzhao Wu, et. al. [21] propose a collaborative learning system for sentiment classification to deal with sparse annotated data of few domains. The paper also proposes to speed up the process by implementing the collaborative learning based on multi-task learning to train sentiment classifiers for multiple domains simultaneously and also by usage of parallel processing.

To deal with scarcity of labelled training data the paper proposes to find out the relatedness between the domains so that related domains can effectively use the less amount of annotated data effectively.

The paper proposes to divide the sentiments into two parts in each domains. The first part would contain the global words, which do not change their polarity across any domains. And, the second which would have domain specific sentiment words. The domain specific captures the domain specific knowledge of the specific domain.

The system also encourages sharing of sentiment information between similar domains. This is done by using the similarities between domains. Also, two kinds of domain similarity measures have been implemented in the system, one based on textual similarity and the other one based on sentiment expressions.

Shaowu Zhang, et. al. [22] propose to achieve cross domain sentiment analysis with the key sentence extraction method. Generally not all part or words of the review are important for sentiment analysis, the paper takes the advantage of this fact to extract only that key sentence which would contain in itself the polarity of the whole document.

Three factors are considered while extraction of the key sentence. First is the Sentiment purity, as a representative of the whole document the purity is the sentiment orientation is very important. The sentence purity has been calculated as the sentence contribution factor. The second is the Keyword property, the paper observes that in human habits people tend to summarise their opinion with key words such as ‘overall’, ‘in my opinion’, etc. These key words helps in proper extraction of key sentence. And the third factor is the position property, as it is generally seen that the introduction and the summary generally contain important sentiment orientation information this is considered for key sentence extraction.

The classifier is trained on labelled training data of two types, key views and details views. Detail views are irrelevant for sentiment analysis. And the paper assumes that the sentiment orientation of the extracted key sentence from any domain will be predicted accurately by the classifier.

The paper proposes a different idea for cross domain sentiment analysis, but the assumption made by the paper that sentiment orientation of key words extracted from any domain will be predicted by the classifier is not entirely true. The accuracy of the sentiment classification will depend of the expanse of the labelled training data.

F. Bisio et.al. [23] identify two issues with regards to the cross domain sentiment analysis, first is with regards to the training of the classifier and the second is with regards the generalization ability of the classifier in cross domain analysis. To solve this issues the paper proposes a distance based predictive model and which tries to combine simplicity and modularity, and delegates domain independent aspects to the definition of semantic-based metrics. The paper claims to provide better results in cross domain sentiment analysis.

Jyoti S. Deshmukh and Amiya Kumar Tripathy [24] propose to weight the training data. This is achieved with the help of the maximum entropy and point wise mutual information method. But compared to paper [13] the proposed system does not provide any new and better solution.

Rui Xia, et. al. [25] propose a feature ensemble plus sample selection (SS-FE) method to deal with cross domain adaptation. The paper identifies two major problems labelling adaptation and instance adaptation. Labelling adaptation deals with the problem of dealing with words having different meanings in different domains. And, instance adaptation deals with the problem of change in the vocabulary or word frequency in different domain. Labelling adaptation method proposed in the paper is based on feature ensemble (FE). This idea is based on the observation of the authors that, features with different type of part-of-speech (POS) tags have a distinct change in distribution in domain adaptation. The paper terms the heavily changing features as *domainspecific*, and the slightly changing features as *domain-independent*. And to deal with the instance adaptation the concept of principal component analysis is used.

Swati Sanagar and Deepa Gupta [26] propose to use iterative Latent Semantic Analysis technique to develop a polarity lexicon. The paper claims that this polarity lexicon is adaptable across multiple target domains. The polarity lexicon is developed by learning the seed words from multiple domains. The paper gives a very simplistic solution to the domain adaptation problem. The major drawback of the proposed system is the requirement of large number of labelled training data, which the make the implementation of the system a complex task.

Kaili Mao, et. al. [27] for performing cross-domain sentiment analysis of Chinese product reviews combine the Lexicon-based and Learn-based techniques (CLL). For this purpose the authors consider lexicons from three different domain domains books, hotels and electronics. And, further the paper considers four feature for each of the domains and build six classifier to achieve cross domain sentiment analysis. The development and training of six classifier is a difficult task that the papers tries to undertake.

Robert Remus [28] propose to achieve domain adaptation by directly selecting instances from source domain for the target domain instances by measuring the domain similarity and domain divergence of the source domain and the target domain. Even though the paper claims to provide a better solution to the cross domain adaptation problem the similarity and variance measurement is a long process which make the proposed system computationally very expensive.

Huimin Wu and Qin Jin [29] try to build on the transfer learning method proposed in paper [15]. The authors are of the opinion that before transfer learning is applied, that is, before the similarity of similar words have been found out appropriate data sample should be selected from the source domain. The paper claims that if this is done then the time complexity of find the similar words can be mitigated to some extent.

Whereas Suman D. Roy, et. al. [30] propose to work with transfer learning on social media. The paper proposes to develop a social transfer technique where transfer learning techniques can be implemented at runtime. For transfer learning at real time the paper tries to solve two main problems, the system must learn the interconnected pattern of shared features between the source and the target domain data, and since the topics modelled from social stream changes with the real world trends, the system requires a transfer framework that allows inclusion of topics on real-time basis. On the other hand Yulan He et. al. [31] proposes to extend the Joint Topic Model proposed in [9] for the social media.

Santosh Kumar, et. al. [32] propose to create clusters by using the unlabelled source and target domain data and then label the clusters with the help of labelled target domain data. Even though the paper claims to achieve cross domain adaptation the major drawback of the proposed system is that this system will work only when the target domain is known and during the training and testing phase a labelled target domain data is used. Also, the paper will fail to achieve sentiment classification.

Cong-Kai Lin, et. al. [33] is of the view that generally in cross domain adaptation the source and the target domains are taken as a whole, but in real world scenario the reviews and opinions on any product are already categorised and organised according to the type of product in a tree type structure. Making use of this information the paper proposes to use a general ensemble algorithm which takes into account the model application, the model weight and the strategies for selecting the most related models with respect to the target domain data. Then to achieve classification of cross domain data sentiment classification technique Support Vector Machine (SVM) and the transfer learning algorithm Spectral Features Alignment (SFA) were applied.

Yang Bao, et. al. [34] proposes a model called Partially Supervised Cross-Collection LDA topic model (PSCCLDA) for cross-domain learning with the purpose of addressing two issues. The paper tries to reduce distributional difference between domains for accurate classifier learning. Also,

domain-independent and domain-specific latent features have been distinguished for proper alignment of the domain-specific features. The paper claim to produce better results across different datasets.

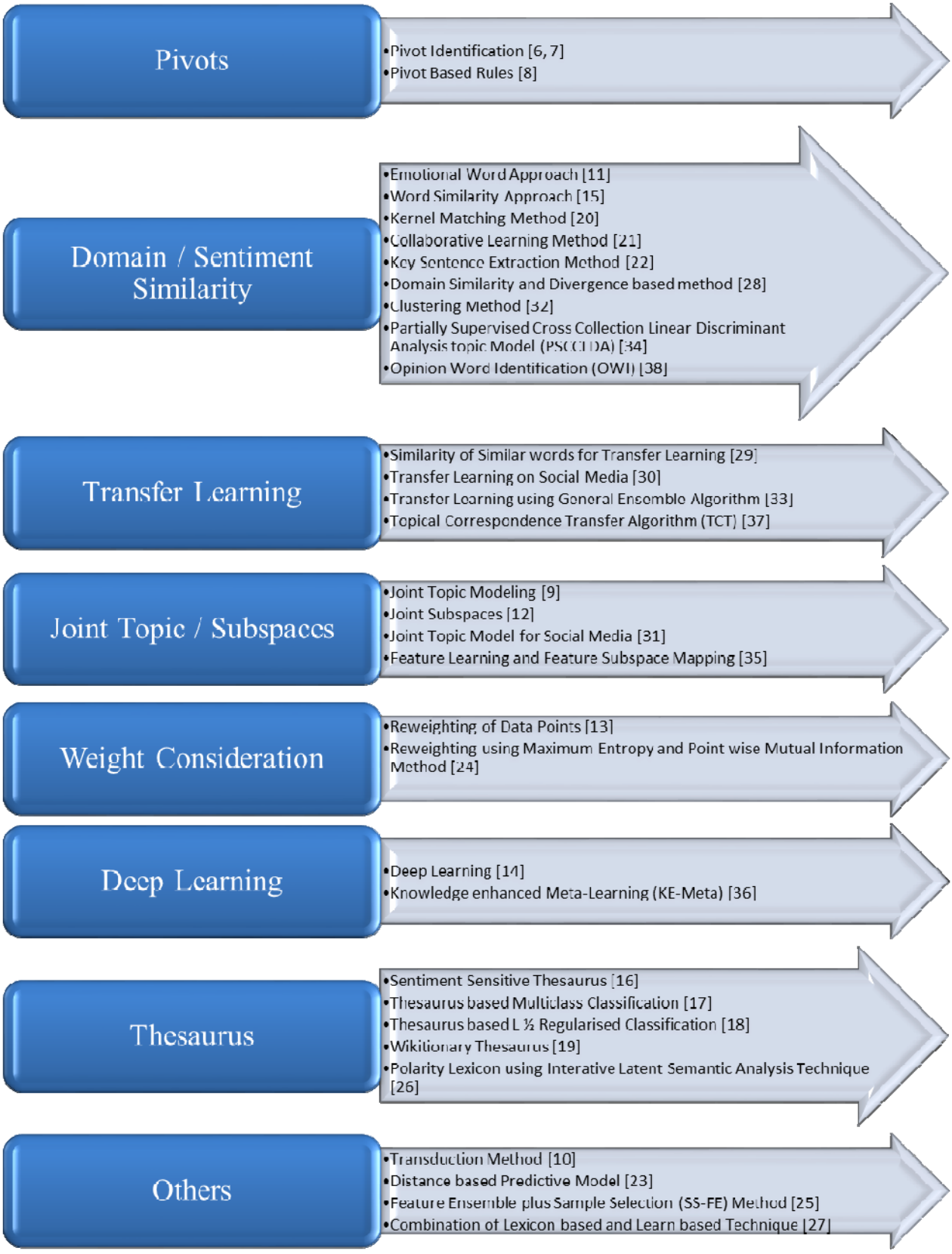

Fig. 1 Cross Domain Sentiment Analysis Techniques

N. X. Bach, et. al. [35] proposes a method that uses feature learning and feature subspace mapping. It uses Word embedding's and canonical correlation analysis (CCA) to address vocabulary mismatches that occurs between source and target domains. Even though the proposed system is easy to adapt and can be used for a range of other natural language processing tasks, the experimental analysis has been performed in a closed environment.

M. Franco-Salvador, et. al. [36] proposes a Knowledge-enhanced Meta-learning (KE-Meta) algorithm that uses BabelNet multilingual semantic network. By summing up different type of classifier the classification of the documents is performed as per its polarity. The extra information provided by the proposed method helps in information gain of base classifier which is better than the n-gram-based classifiers. However domain adaptation is not performed by the KE-Meta algorithm.

G. Zhou et. al. [37] proposes to develop an algorithm called Topical Correspondence Transfer (TCT). It has been proposed that the algorithm will learn domain specific information from different domain and combine them into unified topics, with the help of shared topics across all domains. The paper claims that the proposed system performs better than the existing cross domain sentiment classification.

Y. Tsai et. al. [38] propose to develop a system called "The Opinion Word Identification (OWI)" with the help of active learning scheme called Query-by-Committee (QBC).The QBC helps in improving the system by improving the accuracy by training the system on fewer annotated sentences.

III. ANALYSIS

A brief analysis of the important techniques and papers has been given in the table below. The analysis has been carried out on the basis of three aspects. These include the techniques used or proposed in the paper, dataset used for training and testing purpose, and the claimed cross domain sentiment analysis accuracy.

TABLE I. SUMMARY OF IMPORTANT CROSS DOMAIN SENTIMENT ANALYSIS TECHNIQUES

Sr No.	Paper Title	Technique used	Dataset used	Accuracy
1	Cross-domain Sentiment Classification using Sentiment Sensitive Embeddings [8]	Sentiment Sensitive Embeddings using Pivots mapping	Amazon product reviews : books, DVD, electronics and Kitchen appliances	70.59 %
2	Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach [9]	Joint Topic Modeling using Opinion pairs and aspect terms	Publicly available user-generated review data : Game, CD and Hotel	83.73 %
3	TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets [10]	Transductive learning Method And Twitter Hierarchical mapping	Runtime Tweeter Data	58.54 %
4	Collaboratively training sentiment classifiers for multiple domains [21]	Collaborative training and Domain similarity	Amazon product reviews : books, DVD, electronics and Kitchen appliances	66%
5	Cross-Domain Recognition by Identifying Joint Subspaces of Source Domain and Target Domain [12]	Joint Subspaces Method	Image Datasets: Amazon and Caltech	80.53 %
6	Prediction Reweighting for Domain Adaptation [13]	Data points Reweighting and Label propagation	Amazon product reviews : books, DVD, electronics and Kitchen appliances	79.9%
7	Deep Learning and Ensemble Methods for Domain Adaptation [14]	Deep Learning Method	Amazon product reviews : books, DVD, electronics and Kitchen appliances	85%
8	Transfer Learning by Linking Similar Feature Clusters for Sentiment Classification [15]	Transfer Learning based on Similarity	Amazon product reviews	81.85 %
9	Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus [16]	Sentiment Sensitive Thesaurus	Amazon product reviews : books, DVD, electronics and Kitchen appliances	80.91 %
10	Feature Space Independent Semi-Supervised Domain Adaptation via Kernel Matching [20]	Kernel Matching	Amazon product reviews : books, DVD, electronics and Kitchen appliances	78.86 %

In Fig. 1 all the important cross domain sentiment analysis techniques have been categorized with regards to the basic way through which the researchers propose to achieve cross domain sentiment analysis. These include usage of pivots, finding of domain or sentence similarity, usage of transfer

learning technique, creation of joint topic modelling or subspaces, creation of thesaurus, etc.

Of the different cross domain sentient analysis paper discussed, ten important paper have been analysed in table 1. And the accuracy of cross domain sentiment analysis obtained ranges from 50 to 85%.

If category-wise cross domain sentiment analysis accuracy is seen then the usage of pivots gives an accuracy of around 70%; for domain similarity the accuracy obtained is around 80%; for the creation of subspace the accuracy is around 83%; the reweighting technique produces 80% accuracy and creation of thesaurus produces an accuracy of around 80%. For standard comparison Amazon Product Reviews has been taken as the dataset for training and testing purposes. This dataset contains reviews on books, DVDs, electronics and kitchen appliances.

For industrywide use and acceptance of cross domain sentiment analysis the accuracy of the cross domain sentiment analysis systems would have to be further improved.

IV. CONCLUSION

The increase in the number of products and product categories, and the need for proper sentiment analysis of the user reviews has created multifaceted problems. To solve this problem many methods have been proposed by researchers in a very short period for cross domain sentiment analysis or cross domain adaptation. Some important works include the proposed techniques like Sentiment Sensitive Embeddings using Pivots mapping, Transfer Learning based on Similarity, Data points Reweighting, Joint Subspaces Method, etc.

Even though many important systems has been proposed in the field of cross domain sentiment analysis there is still ample scope of improvement for proper adaptation of the domains. For this better domain adaptation techniques are required which would produce better and more accurate cross domain sentiment analysis results.

Proper and accurate identification of the source domain for a specific target domain is the foundation for achieving more accurate sentiment analysis results. So, further research in this regard is desirable.

For proper identification of the source domain for a specific target domain the feature similarity and divergence plays a very important role. Detailed study is required for understanding which domains will be able to adapt to which other domains.

REFERENCES

[1] Editorial Staff. (2017, Jan 12). *Internet Trends, Stats & Facts in the U.S. and Worldwide 2017* [Online]. Available: http://www.hotel-online.com/press_releases/release /internet-trends-stats-facts-in-the-u.s.-and-worldwide-2017, Last accessed 19th Feb, 2017.

[2] Editorial Staff. (2016, June 8). *How Internet in India Will Look in 2020: 12 Exciting Statistics!* [Online]. Available: http://trak.in/tags/business/2016/06/08/india-internet-growth-2020-statistics/, Last accessed on 19th Feb, 2017.

[3] Neeraj M. (2016, Feb. 8). *Mobile Internet Users In India 2016: 371 Mn by June, 76% Growth In 2015* [Online]. Available: https://dazeinfo.com/2016/02/08/mobile-internet-users-in-india-2016-smartphone-adoption-2015/, Last Accessed on 19th Feb 2017.

[4] Michael Haupt. (2016, May 2). "Data is the New Oil"—A Ludicrous Proposition Natural resources, the question of ownership and the reality of Big Data [Online]. Available: https://medium.com/twenty-one-hundred/data-is-the-new-oil-a-ludicrous-proposition-1d91bba4f294#.3vkhfhik1, Last Accessed on 19th Feb, 2017.

[5] Christine Day. (2015, March 17). *The Importance of Sentiment Analysis in Social Media Analysis* [Online]. Available: https://www.linkedin.com/pulse/importance-sentiment-analysis-social-media-christine-day, Last accessed on 19th Feb, 2017.

[6] John Blitzer, Mark Dredze and Fernando Pereira, "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification", ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, June 23-30, 2007.

[7] John Blitzer, Ryan McDonald, and Fernando Pereira, "Domain adaptation with structural correspondence learning", Empirical Methods in Natural Language Processing (EMNLP), 2006.

[8] Danushka Bollegala, Tingting Mu and John Y. Goulermas, "Cross-domain Sentiment Classification using Sentiment Sensitive Embeddings", IEEE Transactions on Knowledge and Data Engineering, 2015.

[9] Zhen Hai, Gao Cong, Kuiyu Chang, Peng Cheng, and Chunyan Miao, "Analyzing Sentiments in One Go: A Supervised Joint Topic Modeling Approach", IEEE Transactions on Knowledge and Data Engineering, 2017.

[10] Shenghua Liu, Xueqi Cheng, Fuxin Li, and Fangtao Li, "TASC:Topic-Adaptive Sentiment Classification on Dynamic Tweets", IEEE Transactions on Knowledge and Data Engineering, 2014.

[11] Kun-Hu Luo, Zhi-Hong Deng, Liang-Chen Wei, and Hongliang Yu, "JEAM: A Novel Model for Cross-Domain Sentiment Classification Based on Emotion Analysis", Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17-21 September 2015, Association for Computational Linguistics, pages 2503–2508.

[12] Yuewei Lin, Jing Chen, Yu Cao, Youjie Zhou, Lingfeng Zhang, Yuan Yan Tang, and Song Wang, "Cross-Domain Recognition by Identifying Joint Subspaces of Source Domain and Target Domain", IEEE Transactions On Cybernetics, 2016.

[13] Shuang Li, Shiji Song, and Gao Huang, "Prediction Reweighting for Domain Adaptation", IEEE Transactions On Neural Networks And Learning Systems, 2016.

[14] Debora Nozza, Elisabetta Fersini, and Enza Messina, "Deep Learning and Ensemble Methods for Domain Adaptation", IEEE 28th International Conference on Tools with Artificial Intelligence, 2016, pp. 184-189.

[15] Wenjie Zhang, Hui Zhang, Deqing Wang, Rui Liu, He Zhang, Xianlin Jiang, and Yong Chen, "Transfer Learning by Linking Similar Feature Clusters for Sentiment Classification", IEEE 28th International Conference on Tools with Artificial Intelligence, 2016, pp. 1019-1026.

[16] Danushka Bollegala, David Weir and John Carroll, "Cross-Domain Sentiment Classification using a Sentiment Sensitive Thesaurus", IEEE Transactions On Knowledge And Data Engineering.

[17] K. Aarthi, C. S. Kanimozhi Selvi, "Enhancing Accuracy in Cross-Domain Sentiment Classification by using Discounting Factor", International Journal Of Engineering And Computer Science, Volume 3 Issue 5 may, 2014 Page No. 5879-5885.

[18] J. Karthikeyan and E. Suresh, "An Improved Cross-Domain Sentiment Classification using L1/2 Penalty Logistic Regression" International Journal of Engineering Research & Technology (IJERT) Vol. 3 Issue 2, February – 2014.

[19] P.Sanju and T. T. Mirnalinee, "Cross Domain Sentiment Classification Using Enhanced Sentiment Sensitive Thesaurus (ESST)", 2013 Fifth International Conference on Advanced Computing (ICoAC), pp. 370-375.

[20] Min Xiao and Yuhong Guo, “Feature Space Independent Semi-Supervised Domain Adaptation via Kernel Matching”, IEEE Transactions On Pattern Analysis And Machine Intelligence, 2015.

[21] Fangzhao Wu, Zhigang Yuan, and Yongfeng Huang, “Collaboratively Training Sentiment Classifiers for Multiple Domains”, IEEE Transactions on Knowledge and Data Engineering, 2017.

[22] Shaowu Zhang, Huali Liu, Liang Yang, and Hongfei Lin, “A Cross-Domain Sentiment Classification Method Based on Extraction of Key Sentiment Sentence”, NLPCC 2015, Springer International Publishing, Switzerland, 2015, pp. 90–101.

[23] F. Bisio, P. Gastaldo, C. Peretti, R. Zunino, and E. Cambria, “Data intensive review mining for sentiment classification across heterogeneous domains,” Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONAM ’13, pp. 1061–1067.

[24] Jyoti S. Deshmukh and Amiya Kumar Tripathy, “Mining Multi Domain Text Reviews Using Semi-Supervised Approach”, 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th & 18th March 2016, Coimbatore, TN, India.

[25] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria, “Feature Ensemble Plus Sample Selection: Domain Adaptation for Sentiment Classification”, IEEE Intelligent Systems, May/June 2013, pp. 10-18.

[26] Swati Sanagar and Deepa Gupta, “Adaptation of Multi-domain Corpus Learned Seeds and Polarity Lexicon for Sentiment Analysis”, IEEE 2015 Intl. Conference on Computing and Network Communications (CoCoNet'15), Dec. 16-19, 2015, Trivandrum, India, pp. 50-58.

[27] Kaili Mao, Jianwei Niu, Xuejiao Wang, Lei Wang and Meikang Qiu, “Cross-Domain Sentiment Analysis of Product Reviews by Combining Lexicon-based and Learn-based Techniques”, 2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), and 2015 IEEE 12th International Conf on Embedded Software and Systems (ICESS), pp. 351-356.

[28] Robert Remus, “Domain Adaptation Using Domain Similarity- and Domain Complexity-based Instance Selection for Cross-domain Sentiment Analysis”, IEEE 12th International Conference on Data Mining Workshops, 2012, pp. 717-723.

[29] Huimin Wu and Qin Jin, “Improving Emotion Classification on Chinese Microblog Texts with Auxiliary Cross-Domain Data”, IEEE International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, pp. 821-826.

[30] Suman D. Roy, Tao Mei, Wenjun Zeng, and Shipeng Li, “SocialTransfer: Cross-Domain Transfer Learning from Social Streams for Media Applications”, MM’12, October 29–November 2, 2012, Nara, Japan, ACM, pp. 649-658.

[31] Yulan He, Chenghua Lin, Wei Gao, and Kam-Fai Wong, “Dynamic Joint Sentiment-Topic Model”, ACM Transactions on Intelligent Systems and Technology, Vol. 5, No. 1, Article 6, Publication date: December 2013.

[32] Santosh Kumar, Xiaoying Gao, and Ian Welch, “Cluster-than-Label: Semi-supervised Approach for Domain Adaptation”, IEEE 31st International Conference on Advanced Information Networking and Applications, 2017, pp. 704-711.

[33] Cong-Kai Lin, Yang-Yin Lee, Chi-Hsin Yu, Hsin-Hsi Chen, “Exploring Ensemble of Models in Taxonomy-based Cross-Domain Sentiment Classification”, CIKM’14, November 3–7, 2014, Shanghai, China, ACM, pp. 1279-1288.

[34] Yang Bao, Nigel Collier, and Anindya Datta, “A Partially Supervised Cross-Collection Topic Model for Cross-Domain Text Classification”, CIKM’13, October 27 – November 01 2013, San Francisco, CA, USA, ACM, pp. 239-247.

[35] N. X. Bach, V. T. Hai, and T. M. Phuong, “Cross-domain sentiment classification with word embedding’s and canonical correlation analysis,” in Proceedings of the Seventh Symposium on Information and Communication Technology, 2016, pp. 159–166.

[36] M. Franco-Salvador, F. L. Cruz, J. a. Troyano, and P. Rosso, “Cross-domain polarity classification using a knowledge-enhanced meta-classifier,” Knowledge-Based Systems, vol. 86, 2015, pp. 46–56.

[37] G. Zhou, Y. Zhou, X. Guo, X. Tu and T. He, “Cross-domain sentiment classification via topical correspondence transfer,” Neurocomputing, vol. 159, 2015, pp. 298–305.

[38] Y. Tsai, R. T. Tsai, C. Chueh, and S. Chang, “Cross-Domain Opinion Word Identification with Query-By-Committee Active Learning,” Technologies and Applications of Artificial Intelligence, 2014 , pp. 334–343.